Commentary on Isager et al. (2021) Reflections on the Replication Value (RV) and a Proposal for Revision

Bakker, B.N.; Bomm, L.; Peterson, D.J.

DOI
10.31234/osf.io/xgekt

Publication date
2024

Document Version
Submitted manuscript

License
CC BY

Commentary on Isager et al. (2021) Reflections on the Replication Value (RV) and a Proposal for Revision*

Bert N. Bakker, Linda Bomm, & Delaney J. Peterson[1]

[1]Amsterdam School of Communication Research, University of Amsterdam

July 17, 2024

*This commentary is a response to the following paper: Isager, P. M., van 't Veer, A. E., Lakens, D. (2021, August 24). Replication value as a function of citation impact and sample size. https://doi.org/10.31222/osf.io/knjea

Submitted to Meta-Psychology. Participate in open peer review by sending an email to open.peer.reviewer@gmail.com. The full editorial process of all articles under review at Meta-Psychology can be found following this link:
https://tinyurl.com/mp-submissions
You will find this preprint by searching for the first author's name.

Introduction

To help researchers determine what studies to replicate, Isager et al. (2021, p.1) introduced the Replication Value (RV), "a proxy for expected utility gain". In this commentary, we point out that there are scenarios where people can opt to replicate studies with a lower RV compared to those with a higher RV. Methodologically, we highlight that the differences in the RV become very small when comparing studies with different large (e.g., N>500) samples. To adjust for this, we demonstrate that a modification to the RV-equation – by log transforming the sample size – leads to a RV that discriminates between studies with large sample sizes with greater precision.

When to use the RV

In communication research, a field we work in, we see that replication studies are uncommon (Keating & Totzkay, 2019). Yet, when asked about replications, communication researchers are positive about their value (Bakker et al., 2021). As the uptake of replications (hopefully) increases, we think the RV will help scholars to determine what study to replicate. At the same time, we wonder whether the RV in its current form consistently picks up the "right" studies to replicate. Below, we highlight two scenarios where a study with a relatively low RV – compared to other studies – might still be useful to replicate.

First, Chambers (2017), for instance, argued that scholars should build replication into their work. For instance, a first paper of a PhD dissertation could be a direct replication of a study that is informing the dissertation. This could be the study with the highest RV. The PhD student may choose to replicate a study with relatively high uncertainty (i.e., low sample size), because they want to increase the – currently low – confidence in the study's results. But we also see scenarios whereby a PhD student chooses to replicate a study with a lower RV. For instance, that particular study may be more at the centre of the idea that the PhD student wants to build upon. Furthermore, the PhD student may also choose to replicate a study with relatively low uncertainty (i.e., high sample size), because they want to further solidify the – already high – confidence they have in the study's results.

Second, when testing the "repeatability" (Freese & Peterson, 2017) of a study across contexts, the RV might not be the metric to guide decisions. Let us illustrate this. Recently, some of us replicated a study by Soroka et al. (2019) which has an RV of 1,67 (371 citations in 6 years and a sample size of N=1100). This RV is lower than other related studies (e.g., Lang et al., 1996; Soroka & McAdams, 2015). But we chose to replicate this study because of its agenda-setting nature and impact on different fields. To assess the "repeatability" of the findings by Soroka et al. (2019) in a another context (e.g., Freese & Peterson, 2017,

p. 152) – using the same stimuli and the same procedures – we directly replicated and extended the study in the Netherlands (see, Dubèl et al., 2024). In a case where more direct replications are conducted in other contexts, the RV may not be the only decision criterion to judge the relevance of a replication.

To summarize, in a replication project, the RV could help to "select" a study that one wants to replicate. But in literature that is cumulatively generating knowledge or when testing the "repeatability" of a study, the RV might not be apt for guiding decisions.

The RV when choosing between studies with large sample sizes

In this section we ask the following question: how well does the RV capture uncertainty when sample sizes are relatively large? The RV formula that Isager et al. (2021) propose is a function of citation count and sample size. Sample size is a "proxy" of uncertainty. Doing so, the RV is elegant and, relatively, straightforward to calculate. The RV put forward by Isager et al. (2021) is as follows:

$$\text{RV} = \left(\frac{\text{citations}}{\text{years\_since\_pub} + 1}\right) \times \left(\frac{1}{\sqrt{(n)}}\right)$$

where:

- citations is the number of citations the article has received.
- years_since_pub is the number of years since the article was published + 1
- n is the sample size of participants in the study.
- $\sqrt{}$ denotes the square root of the sample size.

Our interest is in the sample size as an expression of uncertainty. Let's take the scenario where a researcher can choose between two studies with relatively small sample sizes: Study A and Study B. Study A and B both have 20 citations in the first year but a different sample size. Study A has a sample size of *N*=50. For the second part of the equation – $\frac{1}{\sqrt{n}}$ – this results in a multiplicative term of 0.141 ($\frac{1}{\sqrt{50}}$). Study B has a sample size of *N*=200 which results in a multiplicative term of 0.071 ($\frac{1}{\sqrt{200}}$). Thereby the RV for Study A is 1.41 ($\left(\frac{20}{1+1}\right) \times 0.141$), while the RV of Study B is .71 ($\left(\frac{20}{1+1}\right) \times 0.71$). It is straightforward to conclude that Study A has a larger RV than Study B and would be the study to replicate.

In many fields, large sample sizes are common. Yet, these fields are also confronted with small population based effect sizes. Therefore, a study with sample size of 500 is much more uncertain than a study with 1000 or 3000 observations (Arel-Bundock et al., 2024). Crucially, when the sample size increases, the value of the second part of the RV equation ($\frac{1}{\sqrt{n}}$) will approximate zero: dividing 1 by a large number will result in a value close to zero. As a consequence, a study's RV will be mostly an expression of its citations, and the larger a study's sample, the more this will be the case. To illustrate the potential overweighting of citations over sample size in the current RV formula, we conducted a simulation whereby we varied the sample size of the study and plotted the RV over the range of the average number of citations (x-axis) and the years since publication (y-axis) – to access the code, see: https://osf.io/kqjpz/. Darker red colours indicate a higher RV and darker blue colours indicate a lower RV. As the top-left panel shows, at a sample size of 50, the RV has a lot of discriminatory power: there are lower and higher RV values. When the sample size increases to 500, 1000 or 3000 observations, we see that the RV approximates zero in almost all cases - as can be seen from the dark blue panels in Figure 1.

Yet, when researchers use the RV to select among studies that are characterized by larger sample sizes, they will have to learn how to interpret small differences in the RV. Isager et al. (2021) acknowledge this when they argue that the RV is relative, stating that the "relative rank-order replication value within the set [of replication studies] matters for answering 'which of the candidates in this set would I increase expected utility the most by replicating?"' (line 180). Our simulation shows that with larger sample sizes, the uncertainty expression receives less weight in the RV and the estimates of the RVs of studies with large sample sizes become more closely related to each other.

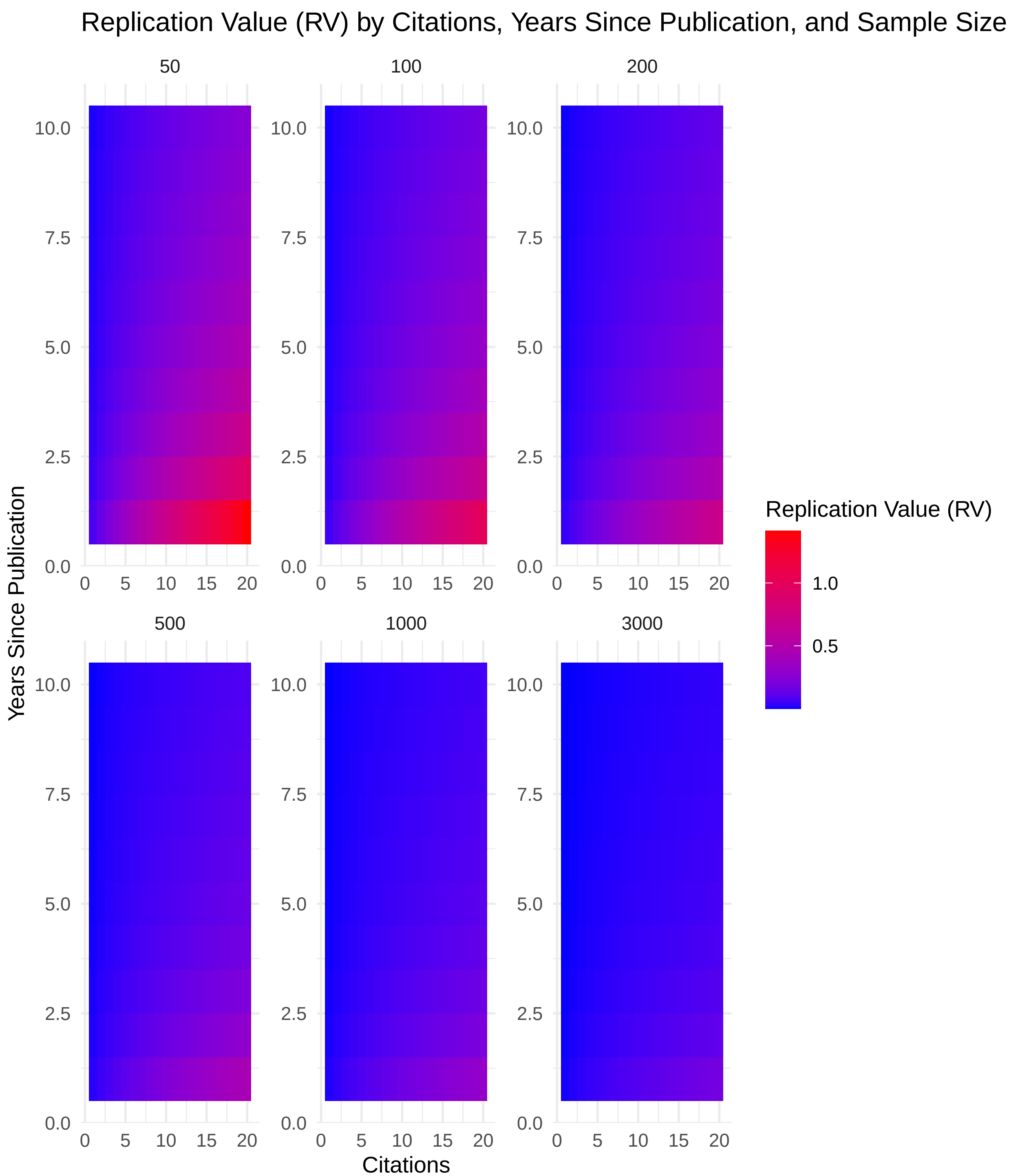
Would it help to modify the RV so that it can better differentiate between a body of studies with a larger sample size? We illustrate what happens with the RV if we use the logarithm of the sample size as a measure of uncertainty (see Figure 2). In doing so, the influence of the sample size on the RV is reduced as it grows. As a consequence, the discriminatory power of the RV is improved at larger sample sizes: there is more variation in the RV. Figure 2 illustrates that as the sample size increases across the panels, we see that more recent papers with more citations (the bottom right corner) get a higher RV than older papers with less citations, as is the case in the original RV, see Figure 1. Contrary to the original RV (see Figure 1), we see that our revised formula results in more variation in the RV, as can be seen by the darker red colours in the panels with the larger sample sizes. Doing so, it becomes easier to rank-order and interpret the RVs of studies with larger sample sizes.

Another potential issue for the RV lies in field-specific norms. The RV – or the log-transformed revised version – does not take into account that some fields typically employ larger sample sizes than others. Yet, the RV's authors acknowledge differences between fields in the expression of citations: Isager et al. (2021) explain that "[...] article citation counts tend to systematically vary between research fields [...]" (line 289), and propose to control this with a "'field-weighted citation impact', [...], in which citations are normalized against the average citation count of articles from the same field of science" (line 295). Especially for instances in which the RV may be used by funders to evaluate replication projects across disciplines, building in a field-specific metric for uncertainty would be welcome. Otherwise, there is a risk that funders will evaluate RV-differences without considering the field-specific differences in sample size.

Isager et al. (2021, p.35) conclude that the RV "could be utilized by any researcher, funder, journal, or other stakeholder who wishes to direct limited resources towards important replication targets." Our simulations show that these stakeholders should be aware that at large sample sizes the differences in the RV in its current form will become very small, and that the rank-order of studies could be based upon RV-differences at the level of the first or second decimal. Furthermore, stakeholders should perhaps consider field-specific sample size in their interpretation of the RV.

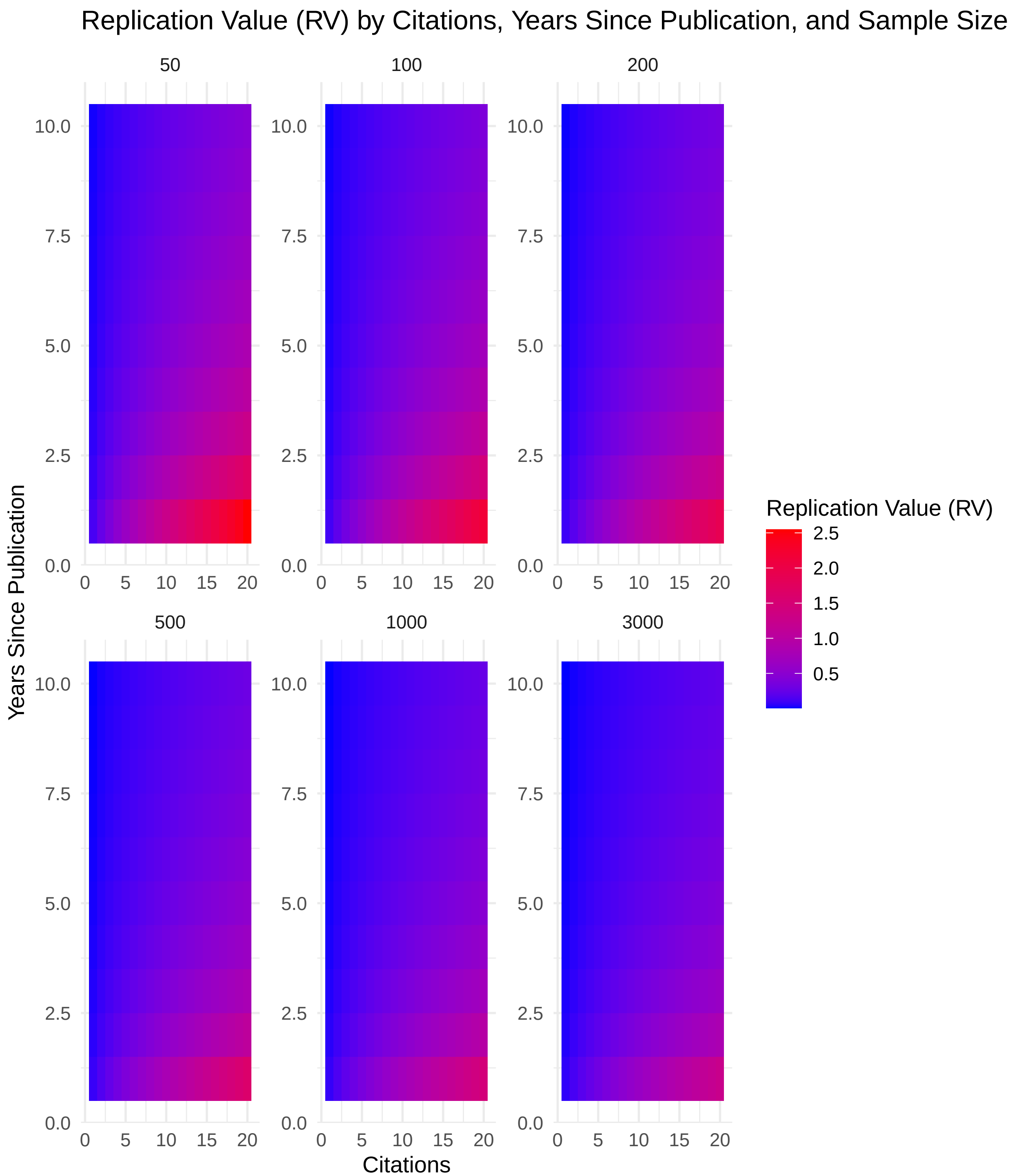
$$\text{RVrevised} = \left(\frac{\text{citations}}{\text{years\_since\_pub} + 1}\right) \times \left(\frac{1}{\log(n+1)}\right)$$

Figure 1: Replication Value over a Range of Different Sample Sizes



Note. Original RV plotted at different sample sizes (panel). In each panel the citations are on the x-axis and the years since publication on the y-axis. Darker red colours indicate a high RV and a darker red colours a lower RV.

Figure 2: Revised Replication Value over a Range of Different Sample Sizes



Note. Revised RV (see equation 2) plotted at different sample sizes (panel). In each panel the citations are on the x-axis and the years since publication on the y-axis. Darker red colours indicate a high RV and a darker red colours a lower RV.

Conclusion

We applaud Isager et al. (2021) for the introduction of the RV. We hope that our comment helps to foster the discussion about the use and functioning of the RV moving forward.

Authorship:

Conceptualization: BNB, LB, DP; Data Curation: BNB; Formal Analysis: BNB; Funding Acquisition: BNB; Investigation: BNB; Methodology: BNB, LB, DP; Project Administration: BNB; Resources: BNB; Supervision: BNB; Validation: BNB, LB, DP; Visualization: BNB; Writing – original draft: BNB, LB, DP; Writing – review & editing: BNB, LB, DP

Conflict of Interest and Funding:

This publication is part of the project "Under pressure: How citizens respond to threats and adopt the attitudes and behaviours to counter them" [project number VI.Vidi.211.055, awarded to Bert N. Bakker] of the NWO Talent Programme VIDI, which is financed by the Dutch Research Council (NWO).
The authors have no conflicts of interests related to this publication.

References

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M., & Stanley, T. D. (2024). Quantitative political science research is greatly underpowered. *The Journal of Politics*.

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, *71*(5), 715–738.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Dubèl, R., Schumacher, G., Homan, M. D., Peterson, D., & Bakker, B. N. (2024). Replicating and extending soroka, fournier, and nir: Negative news increases arousal and negative affect. *Media and Communication*, *12*.

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*(1), 147–165.

Isager, P. M., van't Veer, A., & Lakens, D. (2021). Replication value as a function of citation impact and sample size. *MetaArXiv*. https://doi.org/https://doi.org/10.31222/osf.io/knjea

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239.

Lang, A., Newhagen, J., & Reeves, B. (1996). Negative video as structure: Emotion, attention, capacity, and memory. *Journal of Broadcasting & Electronic Media*, *40*(4), 460–477.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892.

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political communication*, *32*(1), 1–22.